MT for subtitling : User evaluation of post-editing productivity

Koponen, Maarit

European Association for Machine Translation
2020-06-10

Koponen , M , Sulubacak , U , Vitikainen , K & Tiedemann , J 2020 , MT for subtitling : User evaluation of post-editing productivity . in A Martins , H Moniz , S Fumega , B Martins , F Batista , L Coheur , C Parra , I Trancoso , M Turchi , A Bisazza , J Moorkens , A Guerberof , M Nurminen , L Marg & M L Forcada (eds) , Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT 2020) . European Association for Machine Translation , Geneva , pp. 115-124 , Annual Conference of the European Association for Machine Translation , Lisbon , Portugal , 03/11/2020 .

http://hdl.handle.net/10138/320204

MT for subtitling: User evaluation of post-editing productivity

Maarit Koponen * **Umut Sulubacak** * **Kaisa Vitikainen** †* **Jörg Tiedemann** *

* University of Helsinki
{name.surname}@helsinki.fi

† Yle
{name.surname}@yle.fi

Abstract

This paper presents a user evaluation of machine translation and post-editing for TV subtitles. Based on a process study where 12 professional subtitlers translated and post-edited subtitles, we compare effort in terms of task time and number of keystrokes. We also discuss examples of specific subtitling features like condensation, and how these features may have affected the post-editing results. In addition to overall MT quality, segmentation and timing of the subtitles are found to be important issues to be addressed in future work.

1 Introduction

Developments in machine translation (MT) in the last two decades have led to significant improvements in translation quality. The success and popularity of statistical machine translation (SMT) systems were matched and eventually surpassed by neural machine translation (NMT). As quality has improved, the use of MT and post-editing (PE) has also increased in professional translation workflows. Broadly, PE refers to the practice of using MT output as a raw version checked and corrected by the translator. The use of MT and PE has been found to increase productivity in various translation scenarios (e.g. Plitt and Masselot, 2010). However, this workflow appears less common in the field of audiovisual translation (AVT). For example, Bywood et al. (2017) note that while specialised subtitling software with various functionalities are used, technologies like translation memory (TM) or MT have not been widely adopted in AVT. Matusov et al. (2019) suggest that a reason for the lower rate of MT adoption in the AVT field may be that current NMT systems are not suited for the particular features of subtitle translation.

This paper presents a pilot study carried out in November 2019 examining how the use of MT and PE in the subtitling workflow affects the work and productivity of subtitlers. In the study, 12 professional subtitle translators worked on a series of tasks in four language pairs (Finnish→English, Finnish→Swedish, English→Finnish, and Swedish→Finnish). They created interlingual (translated) subtitles for short video clips both with and without MT output. To assess productivity and effort, keylogging data were recorded during these tasks. Task time and technical effort represented by keystrokes were compared between post-editing and translation from scratch.

We first discuss related work on MT for subtitling and approaches to user evaluation of MTPE in Section 2. The MT models and subtitle alignment are presented in Section 3. Section 4 outlines the user data collection, and Section 5 presents the analysis of productivity measures. Section 6 discusses observations on PE changes, followed by future work and conclusions.

2 Related work

2.1 Machine translation for subtitling

Interlingual translated subtitles are a solution (along with dubbing and voice-overs) for bringing movies, television series, documentaries and other video material to audiences who do not understand the original language of the video. Whether dub-

Martins, Moniz, Fumega, Martins, Batista, Coheur, Parra, Trancoso, Turchi, Bisazza, Moorkens, Guerberof, Nurminen, Marg, Forcada (eds.)
Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, p. 115–124
Lisboa, Portugal, November 2020.

bing or subtitling is used varies in different countries and also contexts. Finland, where this study was carried out, is one of the countries where subtitling is predominant for most content types (only children's programming tends to be dubbed).

Subtitling has some features which differentiate it from text translation. Firstly, the source text in subtitling is spoken language, or written representation of spoken language when intralingual subtitles in the language of the original video are used as the source in so-called template translation (e.g. Bywood et al., 2017). The translated subtitles represent source language speech in written target language. Secondly, subtitles have certain technical restrictions related to the number of characters and lines in one subtitle frame, and the length of time the frame is shown on the screen. For example, at the broadcasting company where this study was carried out, subtitle frames contain a maximum of two lines consisting of a maximum of 37 characters, and each frame is on screen from 2 seconds up to 6 seconds. Therefore, subtitle translation commonly involves condensation through solutions like omissions and paraphrases (Pedersen, 2017). Burchardt et al. (2016) also note that issues such as wide variation in subject matter, disfluencies and lack of context in the spoken language as well as the effect of the visual context may present additional challenges for MT.

On the other hand, some authors have suggested that the generally short and relatively simple sentences typical of subtitles would be well-suited for MT. For example, Volk et al. (2010) discuss an SMT system for Swedish→Danish MT of subtitles. In a PE experiment with 6 translators, they report relatively little was edited (average BLEU score between MT and PE for three different TV genres 65.8), with 22% of segments not changed at all. However, no process-based effort measures are reported in that study.

The eTITLE project (Melero et al., 2006) developed a web-based subtitling platform (for English, Spanish, Catalan and Czech) which offered translation memories and MT output from third-party MT engines as a tool for subtitlers. Their tool contains modules for condensation of the machine-translated subtitles and for subtitle placement. Melero et al. (2006) present a user evaluation where one translator translated parts of a movie (English→Czech) either based on the English source text or using MTPE, and report that subtitling the parts with MT was approximately 17% faster than the parts without.

In another study, de Sousa et al. (2011) experimented with MT and TM for DVD subtitling (English→Portuguese). Based on an experiment where 11 volunteers (described as "native speakers of Brazilian Portuguese and fluent speakers of English" with "some experience with translation tasks") alternately translated and post-edited 250 source sentences, de Sousa et al. (2011) report that MTPE was on average 40% faster than translation from scratch.

The SUMAT project (Bywood et al., 2017) developed a cloud-based platform for subtitle translation using MT and post-editing in multiple language pairs, and involved a large-scale user evaluation of productivity and usability of MTPE for subtitling. They collected time data and subjective feedback from 19 professional subtitle translators who translated two files using a source language template, and post-edited MT with and without quality estimation filtering. Bywood et al. (2017) found that MTPE improved productivity (in terms of task time) on average by nearly 40%, although considerable variation was observed in different language pairs and content types. They report the highest increase in English→Dutch (86%) whereas in Spanish→English, a 3.4% decrease of productivity was observed. On average, productivity increased by approximately 14% for scripted vs 50% for unscripted content (Bywood et al., 2017).

Matusov et al. (2019) customised an English→Spanish NMT system for subtitle translation using OpenSubtitles parallel data and other "conversational corpora" like GlobalVoices and TED talks. They report a user experiment where two professional translators subtitled a documentary and a sitcom episode partly from scratch and partly using a source language template and by post-editing two different MT outputs. Based on the experiments, Matusov et al. (2019) estimate average time savings by the translators to be approximately 25% with the customised MT and 5% with the baseline system.

2.2 User evaluation of MT and PE effort

Common approaches to evaluating MT quality include automatic MT metrics such as BLEU (Papineni et al., 2002) or (H)TER (Snover et al., 2006), which calculate similarity scores or edit rates based on the overlap of words or n-grams

between an MT hypothesis and one or more reference translations. These metrics are sometimes used to compare MT output and post-edited versions of the MT as representation of PE effort in terms of the number of words changed during PE (e.g. Volk et al., 2010). However, this product-based approach cannot fully capture the actual effort involved in the PE process. For a more accurate picture of the feasibility of using MTPE, evaluations need to address PE effort in terms of time, technical effort required carrying out for corrections, as well as cognitive effort required for identifying errors and deciding what actions are needed (see Krings, 2001).

Temporal effort can be measured by recording task times (e.g. to the nearest minute) and comparing different types of tasks, such as MTPE versus translation "from scratch" (without MT output), or PE of different MT outputs. More fine-grained time data can be collected using keystroke logging tools like Inputlog (Leijten and Van Waes, 2013), which also provide information about the technical effort involved. Cognitive effort is the most difficult of the three to capture. Approaches to measuring cognitive effort include examining pauses in keylogging, introspective methods, and eyetracking. For an overview of process methodologies, see e.g. Saldanha and O'Brien (2013).

Like the previous studies on MT for subtitling in Section 2.1, the user evaluation reported in this paper addresses productivity in MTPE compared to translation from scratch. However, where prior work has mainly focused on task time or throughput (words or subtitles translated per time unit), we also examine technical effort through keylogging. Effort measures (task time, number of keystrokes) were analysed comparing subtitling from scratch and MT post-editing (see Section 4).

3 Automatic subtitle translation

3.1 Datasets and MT models

For the assessment of MT in subtitle translation, we created sentence-level and document-level translation models from all the parallel data available in OPUS.[1] For Finnish↔Swedish, this includes a bit over 30 million training examples,[2] and for Finnish↔English, roughly 44 million.[3] The training data comes from diverse backgrounds, with sources ranging from Bible translations to software localisation data, official EU publications, and data mined from unrestricted web crawls.

The largest portion of training data is a collection of movie and TV show subtitles derived from the OpenSubtitles (v2018) dataset. For Finnish↔Swedish, this collection contains over 15 million translation units, and for Finnish↔English, it contains almost 30 million translation units. Even though this sub-corpus is quite noisy as well, it fits the task rather well, and we can therefore expect that our models should have a decent performance in the subtitle translation task even without further fine-tuning.

The models we trained rely on the Transformer architecture (Vaswani et al., 2017), the current state of the art in NMT. We apply the implementation from the MarianNMT toolkit (Junczys-Dowmunt et al., 2018), which offers fast training and decoding with the latest features of production-ready NMT. We use the common settings of a multi-layer transformer, with 6 layers on both the encoder and the decoder, and 8 attention heads in each layer. We enable label smoothing and dropout, and use tied embeddings with a shared vocabulary, basically following the recommendations for training transformer models in the MarianNMT documentation. For text segmentation, we apply SentencePiece (Kudo and Richardson, 2018) with models that are trained independently for source and target languages for a vocabulary size of 32,000 in each language. We do not apply any further pre-processing to keep the setup as general as possible, apart from some basic normalisation of Unicode punctuation characters, and parallel corpus filtering using standard scripts from the Moses SMT package (Koehn et al., 2007).

For the document-level models, we apply the concatenative models proposed by Tiedemann and Scherrer (2017) and Junczys-Dowmunt (2019) using units of a maximum length of 100 tokens. Note that sentences and sentence fragments in subtitles are typically very short, and 100 tokens typically cover substantial amounts of context beyond sentence boundaries. We mark sentence bound-

[1] http://opus.nlpl.eu

[2] OPUS corpora used: bible-uedin, DGT, EMEA, EUbookshop, EUconst, Europarl, Finlex, fiskmo, GNOME, infopankki, JRC-Acquis, KDE4, MultiParaCrawl, OpenSubtitles, PHP, QED, Tatoeba, TildeMODEL, Ubuntu, wikimedia

[3] OPUS corpora used: bible-uedin, Books, DGT, ECB, EMEA, EUbookshop, EUconst, Europarl, GNOME, infopankki, JRC-Acquis, KDE4, OpenSubtitles, ParaCrawl, PHP, QED, Tatoeba, TildeMODEL, Ubuntu

aries with special tokens, chunking the training and test data sequentially from the beginning to the end without any overlaps. This procedure creates roughly 3.3 million pseudo-documents for Finnish↔Swedish and 4.7 million documents for Finnish↔English. This means that we have on average about 9 sentences per document, which are concatenated into one long string with boundary markers between sentences.

During test time, we proceed in the same way, creating pseudo-documents from the original input by concatenating subsequent sentences and splitting when a segment exceeds 100 tokens. Sentence-level models are translated in the usual way. In order to examine the translation quality, we applied our models to a dedicated test set taken from a larger set of subtitles from public broadcasts with audio in Finnish, Swedish or English. Intralingual subtitles in the language of the original audio were aligned with interlingual subtitles of the same programme in one of the other two languages. However, it should be noted that the interlingual subtitles are not direct translations of the intralingual subtitles as such. The alignment of subtitle segments in the test set was manually checked and non-corresponding segments were removed. The Finnish and Swedish parts of the dataset also contain intralingual subtitles for the deaf or hard-of-hearing, which were separated in the test set as their own subsets.

The translation results are shown in Table 1, where scores are listed separately for different subsets. Note that the document-level results need to be treated in a special way as they do not automatically match the sentence-level reference translations even when splitting on generated sentence boundary markers. To ensure that the reference and the system output correspond to each other, we apply a standard sentence alignment algorithm implemented in the hunalign package (Varga et al., 2005). We use the re-alignment flag to enable lexical matching as well, which is very beneficial in this monolingual alignment task. BLEU scores may have been negatively affected by this procedure as this alignment is not perfect.

Overall, the results indicate that document-level models seem to be beneficial in the subtitle translation case. The automatic evaluation scores consistently show an improvement over the corresponding sentence-level models for both language pairs and in all directions. However, this encouraging result unfortunately does not carry over to the manual assessment (see Section 5). A reason for this may be at least partially related to the problem of segmentation and time frame alignment, which we introduce below.

benchmark	sentence-level		document-level	
	BLEU	chrF_2	BLEU	chrF_2
fi→sv	18.8	0.443	19.3	0.451
sv→fi	15.7	0.449	16.8	0.462
fi→en	21.5	0.458	23.6	0.472
en→fi	16.0	0.444	17.1	0.454

Table 1: Comparison of BLEU and chrF_2 scores on the benchmark test set for the sentence-level and document-level systems in the language pairs Finnish→Swedish, Swedish→Finnish, Finnish→English, and English→Finnish.

3.2 Subtitle frame alignment

In both sentence-level and document-level translation, we have to treat the results in a way that maps the translations back into the time slots allocated for the original subtitles. Those time slots may include more than one sentence, and sentences may stretch over multiple time slots. Because our translation models are trained on sentence-aligned data, we need to extract sentences first from subtitles, too. We do this using the techniques proposed by Tiedemann (2008), which were also applied to the OpenSubtitles corpus in our training data.

Subtitles converted to sentence-level segments in XML:

```
  <s id="13">
    <time id="T16S" value="00:01:05,960" />
We have to make readmission agreements with other countries, -
    <time id="T16E" value="00:01:12,360" />
    <time id="T17S" value="00:01:12,440" />
so that they would be willing.
  </s>
  <s id="14">
We have to cooperate closely.
    <time id="T17E" value="00:01:17,440" />
  </s>
```

Mapped back to subtitle frames after translation:

```
16
00:01:05,960 --> 00:01:12,360
Meidän on tehtävä
takaisinottosopimuksia muiden maiden kanssa,

17
00:01:12,440 --> 00:01:17,440
jotta ne olisivat halukkaita.
Meidän on tehtävä tiivistä yhteistyötä.
```

Figure 1: Pre- and post-processing of subtitle data before and after translation. Sentences may run over several subtitle frames and multiple sentences and sentence fragments can also appear in the same time frame. The translation comes from a document-level model.

Mapping back to subtitle frames and their time allocations is implemented as another alignment algorithm. We apply a simple length-based al-

gorithm for this, assuming that there is a strong length correlation between the source- and target-language subtitles. The difference to traditional sentence alignment is that we are now only interested in 1-to-n alignments, meaning that each existing subtitle frame in the original input should be filled with one or more segments from the translation. The segments on the target side that we consider are clauses from the generated sentences. For simplicity, we split on any punctuation in the output that is followed by space to approximate the structural segmentation. We then apply the traditional Gale & Church algorithm (Gale and Church, 1993) to optimise the global alignment between source segments (original subtitle frame data) and target segments. For this, we adjust the parameters of the algorithm in two ways: (i) we remove priors and apply a uniform distribution over possible alignment types, and (ii) we change the set of alignment types to include all possible mappings from one source segment to a maximum of four target segments. The mapping between source and target is then created using the original algorithm that ensures a globally optimal mapping according to the model (see Figure 1 for an example). Furthermore, we apply simple heuristics to insert line breaks in order to make subtitles conform to length and formatting constraints. The implementation of the entire procedure is available as an open source package[4].

4 User PE data collection

The subtitling tasks for productivity data collection were carried out in November 2019 at the premises of the Finnish Broadcasting Company Yle. In total 12 translators (3 per language pair) participated in the tasks: 8 in-house translators and 4 freelancers with experience of working for Yle. The participants have between 4 and 30 years of professional subtitling experience in their language pair. Only 2 stated they had previously used MT for subtitling, and 7 others had used MT for other purposes.

The subtitling tasks were carried out using the subtitlers' preferred software (Wincaps Q4 or Spot). To replicate their normal working environment, an external monitor and keyboard were provided, and they had access to the internet as well as terminology and other resources normally used in their work. Process data were logged using Inputlog (Leijten and Van Waes, 2013), which records all keyboard and mouse activity. Windows 10 screen recording software was used to capture video to support the analysis. Pre- and post-task questionnaires were used to collect background information and participants' subjective assessment of the MT output and PE experience. After the tasks, a brief semi-structured interview was also carried out to collect more detailed feedback regarding problems in the workflow and the participants' views on potential improvements. In this paper, we focus on an analysis of the process data.

Subtitling tasks were carried out in 4 language pairs: Finnish→English, Finnish→Swedish, English→Finnish, and Swedish→Finnish. For each source language, six clips were selected from a dataset provided by Yle. Three clips were selected from unscripted European election debates, and three clips from semi-scripted lifestyle or cultural programmes. The individual clips were selected so that each clip (i) forms a coherent, self-contained section of the programme, (ii) is approximately 3 minutes long, and (iii) contains 30–35 subtitle segments.

Each participant completed a total of six tasks where they subtitled two clips "from scratch" without MT output, two clips using output from a sentence-level MT system, and two clips using output from a document-level MT system. The clips and MT outputs were rotated in a round-robin format so that each clip was subtitled once in each condition (no MT output, sentence-level MT output, document-level MT output) by a different participant. Task order was also varied to minimise facilitation effect. The participants were instructed to produce subtitles that would be acceptable for broadcasting, and to use the resources they normally would for their work, but to not spend excessive time in "polishing" any given wording or researching information. No explicit time limit was given for each task, rather, the participants were instructed to work at their own pace.

In the from scratch condition, the participants also created the segmentation and timing of the subtitles following their normal work process. Subtitling templates are not used by Yle for these content types. In the MTPE condition, the participants worked with output that was pre-segmented and timed based on the intralingual subtitles used as source text for the MT (see Section 3.2).

To assess productivity, the process logs were analysed using Inputlog's analysis functions. The

[4] https://github.com/Helsinki-NLP/subalign

task time and the number of keystrokes logged were used as productivity measures. Using Inputlog filters, we focused only on task time and keystrokes in the subtitling software, excluding other activity such as internet searches for terminology or other information. Based on the final subtitles produced, edit rate between the MT output and the final versions were calculated using HTER (Snover et al., 2006) and characTER (Wang et al., 2016). As PE of the subtitles involved also changes to the segmentation, e.g. adding or deleting frames and moving words between frames, subtitle segmentation was ignored and edit rates were calculated as document-level scores to focus on edits affecting the textual content. These measures were then compared between the tasks of creating interlingual subtitles from scratch and MTPE, as well as between PE of the sentence-level and document-level MT outputs described in Section 3.1.

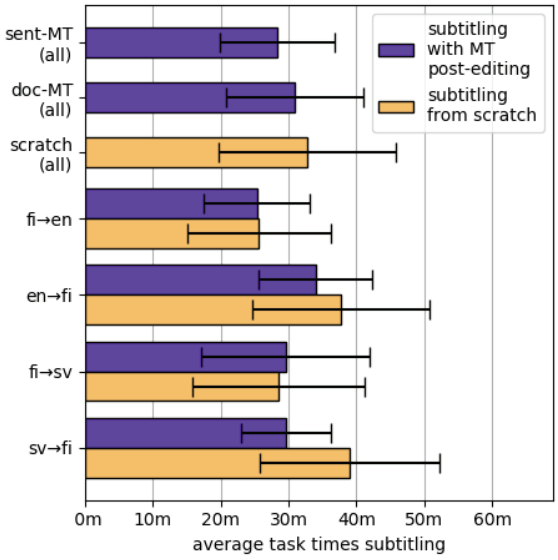

Figure 2: Average task times subtitling through post-editing and from scratch. The top three bars show averages for post-editing sentence- and document-level MT, and subtitling from scratch. The bottom pairs of bars are averages for each language pair. Error bars indicate standard deviation.

5 Comparison of subtitling productivity

Figure 2 shows a comparison of the average subtitling task time for subtitling from scratch and subtitling with MTPE. The topmost three bars show averages for post-editing the sentence- and document-level MT output and for translation from scratch across all language pairs, while the bottom pairs of bars show averages for PE (either MT output) compared to from scratch. On average, post-editing machine-translated subtitles (regardless of MT output) was slightly faster than creating subtitles from scratch. Some differences can be seen between the language pairs: the largest difference in task times is seen in Swedish→Finnish, while the task times for Finnish→English and Finnish→Swedish are nearly equal. No clear difference could be observed between the two different MT outputs, although on average post-editing the sentence-level MT output appeared to be slightly faster.

Figure 3 shows a comparison of technical effort in terms of the average number of keystrokes used when producing subtitles. The topmost three bars show averages for post-editing the sentence- and document-level MT output and for translation from scratch across all language pairs, while the bottom pairs of bars show averages for PE (either MT output) compared to from scratch. On average, post-editing machine-translated subtitles (regardless of MT output) involved fewer keystrokes than creating subtitles from scratch. The reduction in the number of keystrokes is more pronounced than in the case of task times, and seen in all language pairs. Again, no clear difference could be observed between the two different MT outputs, although on average post-editing the sentence-level MT output appeared to involve slightly less technical effort.

Although a detailed analysis of the types of keystrokes is not within the scope of this paper, some observations can be made regarding the distribution of keystroke types. Intuitively, PE reduced the need for text producing keystrokes on average by 54% compared to from scratch, as the MT output provides some of the text needed. However, the number of text deleting keystrokes was 24% higher in PE, as correcting the output also involves removing words or characters. In the from scratch case, the participants needed to create and set the timing for each subtitle frame themselves, which requires keystrokes and/or mouse clicks. In MTPE, the MT output was already segmented and timed based on the intralingual subtitles used as source text, which reduced the associated keystrokes by approximately 32%, but the number of keystrokes shows that the participants found it necessary to change both the segmentation and timing. Changes to subtitle segmentation are discussed in more detail below.

To examine the number of changes between

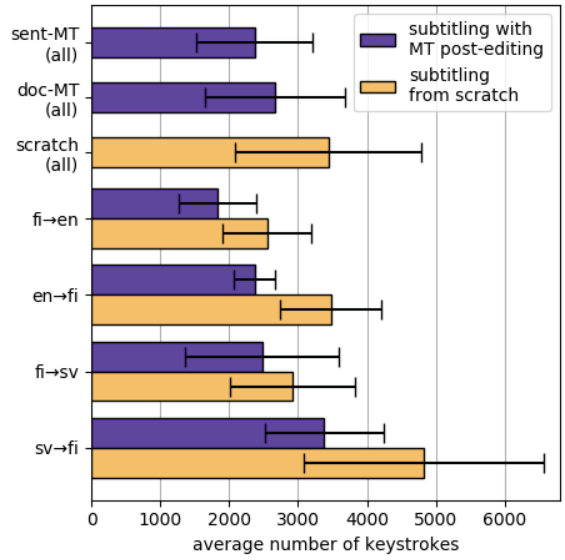

Figure 3: Average numbers of keystrokes subtitling through post-editing and from scratch. The top three bars show averages for post-editing sentence- and document-level MT, and subtitling from scratch. The bottom pairs of bars are averages for each language pair. Error bars indicate standard deviation.

the MT outputs and final PE versions, edit rates were calculated using word-based HTER and character-level characTER. Table 2 shows the HTER and characTER scores for the sentence-level and document-level MT across all four language pairs and for each language pair. The high edit rates (overall average HTER 57.7 and characTER 46.8) indicate considerable rewriting during PE, particularly in the case of English→Finnish. The high HTER score in this language pair may be due to the fact that word-based metrics do not distinguish changed words and changed word forms, which are common in morphologically-rich target languages like Finnish. The considerable difference in the characTER and HTER scores in English→Finnish suggests word form edits are indeed more common in this language pair. However, a similar effect is not seen in Swedish→Finnish. A preliminary analysis of the edits indicates that the participants working on this language pair have added words more frequently than participants in other language pairs. Corresponding to the process metrics, average edit rate for the sentence-level MT output is slightly lower than for the document-level MT. At least partly, this may be explained by the observation that repetition of words or phrases was more common in the document-level MT output.

In addition to the textual content of the MT subtitles, the participants edited both the segmentation of that content into subtitle frames and timing of the frames. On average, the participants increased the number of subtitle frames in the clips by 7% by splitting or adding frames. This tendency was particularly noticeable in Swedish→Finnish (+19%). English→Finnish was the only language pair where the participants reduced the number of subtitle frames (–4%) for example by joining and condensing the textual content of the frames. Comparing the timestamps of the original subtitle frames used for the MT output and the frames in the post-edited files, we observed that only 24% of the original timed frames had been retained in PE. For 27% of frames, either the in or out time had been changed, and for 49% both in and out time were changed.

	HTER	characTER
sent-level	55.1 ± 17.7	45.0 ± 12.3
doc-level	60.3 ± 16.1	48.7 ± 11.1
fi→en	45.6 ± 17.7	39.3 ± 13.5
en→fi	74.1 ± 12.7	48.9 ± 6.4
fi→sv	52.7 ± 13.2	44.1 ± 11.4
sv→fi	58.4 ± 9.5	55.1 ± 9.2
overall	**57.7 ± 16.9**	**46.8 ± 11.8**

Table 2: Comparison of word-level (HTER) and character-level (characTER) edit rates divided by MT system (sentence-level vs document-level) and language pair (Finnish→English, English→Finnish, Finnish→Swedish, Swedish→Finnish).

The intralingual subtitles used as source text were not translated as isolated subtitle frames but rather as sentences or longer passages and then aligned back to the frames (see Section 3.2). However, the heuristics used for alignment were not always successful. In some cases, splitting a segment due to punctuation caused the next segment to become too long and started to push content into the following frames, causing the subtitles to fall out of sync with the audio. Similar issues were also observed due to repetition in the MT output. It is also possible that the sync issues arising from incorrect segmentation may have lead the participants to also change the timing of subtitle frames.

6 Discussion of PE changes

Considerable variation in task times and numbers of keystrokes was observed between different participants. Productivity gains were most evident for participants with the longest average task times

overall. However, 5 out of the 12 participants were in fact slower in PE. Two of them also used slightly more keystrokes, but three were slower despite using fewer keystrokes in PE. These findings are similar to other process studies both on subtitling and other text types (e.g. Plitt and Masselot, 2010; Bywood et al., 2017) showing that potential productivity gains from MTPE vary, and that participants who are already fast benefit less. Fewer keystrokes not necessarily leading to time saving has also been observed in other studies. While the number of keystrokes reflects the technical effort needed, it does not capture the amount of cognitive effort involved in recognising potential errors and deciding on necessary changes.

The edit rates of different participants also vary. At the level of individual subtitlers, average HTER scores range from 31.9 (Finnish→English, participant C) to 84.8 (English→Finnish, participant C). These edit rates are comparable to the HTER scores reported by Matusov et al. (2019) for different MT system outputs, genres and post-editors, which range from 27.8 to 82.7. In our study, the two participants with the highest average edit rates both worked on English→Finnish, and the two with the lowest average edit rates on Finnish→English, but differences are also evident within the same language pair. Since the participants post-edited different MT versions, some variation may be explained by different output quality, but to some extent these differences may also reflect individual preferences. Qualitative observations suggest that while some edits relate to clear MT errors, many are also caused by what appear to be preferential edits; for example, in the Finnish→English clips, one participant accepts the translation "financial discipline" for the Finnish *talouskuri* while another replaces it with "austerity".

A possible factor affecting both productivity and number of changes is PE experience. The participants in this study had little prior experience with MT specifically for subtitling. The subtitlers' productivity and approach to the task may therefore have been affected by the fact that PE was unfamiliar and different from their normal work processes. As Bywood et al. (2017) also note, psychological factors such as unfamiliarity and irritation with MT errors influence productivity. These factors may have also led to preferential and possibly unnecessary changes. More practice working with MT output and pre-segmented subtitles may affect their approaches, e.g. by reducing preferential changes, and increase productivity in this task.

As noted in Section 2.1, the spoken content of the videos and subtitles as a written representation of spoken language differ from each other. Due to technical restrictions, condensation is common in subtitle translation, and may affect the edit rate to some extent. On the other hand, because the source text for the subtitlers consists of not only the written subtitles, but also the audiovisual context, they may make changes based on information in the audio or video of the clip being subtitled.

An example of condensation through omission and paraphrasing can be seen in Table 3, where the participant has combined two subtitle frames (0001 and 0002) in the intralingual subtitles and the MT. This type of condensation was observed particularly in English→Finnish, where the participants reduced the number of subtitle frames.

In contrast to condensation, the participants sometimes added content to subtitles. While some additions correspond to missing words in the MT output, others in fact involve content not present in the intralingual subtitles used as source text for MT. The intralingual subtitles themselves already involve some condensation and paraphrasing, and therefore do not match exactly the spoken audio. Particularly in the Swedish "lifestyle" clips, the intralingual subtitles appear to have been very condensed, and the participants post-editing Swedish→Finnish added both textual content and new subtitle frames. These additions show one effect of the multimodal context: having the omitted information present in the audio led the participants to make additions that would have been unlikely or impossible if only the written subtitles had been available.

Subtitle translators are also affected by the visual context of the video. Changes related to the visual context occur, for example, when the subtitler chooses to replace a pronoun with the referent seen in the video. An example of this appears in one of the Swedish→Finnish clips involving cooking. The expression *de ska kokas mjuka* 'they should be cooked soft' in the dialogue is correctly translated in both MT outputs using the Finnish pronoun *ne* 'they'. However, both participants post-editing MT output for this clip replaced the pronoun with *hedelmät* 'fruit', referring to the fruit being cooked.

Source	MT output (doc)	Post-edited
0001 00:00:00:00 00:00:02:24 Viikonloppuna on vaalitarkkailijoita -	0001 00:00:00:00 00:00:02:24 There will be election observers this weekend -	0001 00:00:00:00 00:00:04:17 There are more election observers there than ever before.
0002 00:00:00:00 00:00:02:24 enemmän kuin ehkä missään muissa vaaleissa	0002 00:00:00:00 00:00:02:24 more than there may be in any other election.	

Table 3: An example of condensation of subtitle content by a post-editor.

These observations suggest that not all changes during PE correspond to MT errors. However, a detailed analysis of the MT outputs and changes carried out during PE would be needed to establish to what extent changes relate to MT errors, subtitling features like condensation, or preferential edits.

7 Future work

Based on the experiment and user feedback, segmentation of the interlingual subtitle content into appropriate chunks is an important issue to be addressed, although using subtitle timing from pre-existing intralingual subtitles was to some extent useful. Potential directions for improving segmentation and timing could involve the use of time information to split the data into coherent blocks separated by significant breaks, and the integration of speaker information into the translation engines to segment subtitles into dialogue turns by leveraging speaker labels or diarisation output. Multimodality can also play a crucial role in segmentation as visual and auditory cues may help in improving the division of verbal content into discourse units. We plan to implement an end-to-end system for subtitle translation and segmentation after Matusov et al. (2019), and investigate how well such a system could generate organic subtitles.

Multimodality may also be useful in optimising translation quality. Augmenting subtitles with information from the visual and auditory modalities could help improve translation accuracy in general. For example, visual information could be helpful in resolving ambiguity. In future work, we will explore incorporating multimodal features in translation in connection with non-linguistic context for language grounding and disambiguation.

A more detailed manual analysis of the types of PE changes made by the participants and their potential explanations (MT errors, subtitling conventions, or preferential changes) is currently underway. Feedback collected from the participants is also being analysed for information regarding the user experience. A second round of user evaluations is also planned for 2020 to collect further data and assess the effect of the new developments of our MT approaches, and to give the participants more experience with post-editing subtitles.

8 Conclusion

This paper presented a user evaluation pilot study of MT and post-editing for subtitles. Based on an analysis of process data collected from 12 professional subtitlers in four language pairs, we presented a comparison of productivity in terms of task time and number of keystrokes when post-editing MT subtitles vs translating from scratch. On average, our results indicate MTPE to be slightly faster and to involve fewer keystrokes than subtitling from scratch. However, considerable variation was observed between different language pairs and participants. We also discussed examples of specific subtitling features like condensation, and how these features may have affected the post-editing results. In addition to overall MT quality, the segmentation and the timing of the subtitles were found to be important issues to be addressed in future work.

Acknowledgments

This work is part of the MeMAD project, funded by the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No 780069).

References

Burchardt, A., Lommel, A., Bywood, L., Harris, K., and Popović, M. (2016). Machine translation quality in an audiovisual context. *Target*, 28(2):206–221.

Bywood, L., Georgakopoulou, P., and Etchegoyhen, T. (2017). Embracing the threat: machine

translation as a solution for subtitling. *Perspectives: Studies in Translatology*, 25(3):492–508.

de Sousa, S. C., Aziz, W., and Specia, L. (2011). Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In *Proceedings of RANLP 2011*, pages 97–103.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Junczys-Dowmunt, M. (2019). Microsoft Translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation*, pages 225–233.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Koehn, P., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., and Moran, C. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th ACL*.

Krings, H. P. (2001). *Repairing texts: Empirical investigations of machine translation post-editing process*. The Kent State University Press, Kent, OH.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 EMNLP*, pages 66–71.

Leijten, M. and Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, 30(3):358–392.

Matusov, E., Wilken, P., and Georgakopoulou, Y. (2019). Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation*, pages 82–93.

Melero, M., Oliver, A., and Badia, T. (2006). Automatic multilingual subtitling in the eTITLE project. In *Proceedings of Translating and the Computer 28*, pages 1–18.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th ACL*, pages 311–318.

Pedersen, J. (2017). The FAR model: assessing quality in interlingual subtitling. *The Journal of Specialised Translation*, 28:210–229.

Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Saldanha, G. and O'Brien, S. (2013). *Research Methodologies in Translation Studies*. Routledge, London and New York.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA 2006*.

Tiedemann, J. (2008). Synchronizing translated movie subtitles. In *Proceedings of LREC'08*.

Tiedemann, J. and Scherrer, Y. (2017). Neural machine translation with extended context. In *Proceedings of the Third DiscoMT*, pages 82–92.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*, pages 590–596.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Volk, M., Sennrich, R., Hardmeier, C., and Tidström, F. (2010). Machine translation of TV subtitles for large scale production. In *Proceedings of the Second Joint EM+/CNGL Workshop*, pages 53–62.

Wang, W., Peter, J.-T., Rosendahl, H., and Ney, H. (2016). CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation*, pages 505–510.